R06525087 郭漢遜

HW2 Video Caption

＊Model description

因為這次作業還沒有完成，感覺寫報告實在沒甚麼說服力，雖然在可能來不及完成BASELINE等級的模型的時候寫報告，還是盡力將這次嘗試過的方法記錄下來・

我目前採用的模型架構是：

輸入 → Bi-LSTM (units = 128) → RepeatVector(次數等於輸出長度) → LSTM (units = 128) → Dense(units = one-hot長度) → 輸出

這個模型感覺就是特別不會WORK，中間也嘗試過其他比較正常的方式，不過不知道為什麼都TRAIN不起來，這個反而至少可以認出前幾個字

比如像是　A man is playing ... , A woman is slicing...

其他的完全不會動！超爛！

輸入就是80 x 4096，輸出的話，稍微檢查一下樣本，只保留了１２個字長以內的樣本，會這樣想是因為，如果短句都講不出來，更不可能講出長的句子吧，言多必失！而且這樣大約還有２萬１千句也保留了大部分的樣本・

不到１２字長的（平均長度７點多）就用"＜PAD＞"補滿・在訓練的時候設定"＜PAD＞"不回傳損失・

＊Attention mechanism

我實作ATTENTION的方式是，在ENCODER並RepeatVector之後，想要讓之後的每一個time step輸出時可以注重在同一個輸入(RepeatVector)的不同部分・

因此要讓每一個Vector都elementwise乘上一個等長但是不同權重分布的Vector・

我的步驟是，假設我想要的輸出長度是 n(個字)，我就把第一層LSTM(units=256)完的結果Repeat n 次，於是得到的shape是（n , 256）・

想把這個結果製作attention的話，先把維度倒過來變成（256, n），於是就好像有了每個小特徵在n個時間步的數值大小組合，然後經過一個全連接層(units = n ,

activation = softmax)學習，原來在n個時間都一樣的東西(因為是Repeat來的)可以透過softmax分配成不同大小，造成每個小特徵只會在某幾個時間步被強調．然後再把維度倒回來．

上述得到的結果，再與Repeat完的原始結果相乘，原本相同的ｎ個Vector，每個的不同部分就有了權重加成變化．進入下一層ＬＳＴＭ．

也有嘗試過不使用單純的全連接而是用LSTM後SOFTMAX相乘，效果一樣不好就是了．

其實也不知道這樣做Attention可不可行，只是直觀上的想法，不過既然沒有成功，我想應該真正的Attention是有其他的做法的．

而我的作法在我的模型上並沒有甚麼效用．或是略為有效，但是因為其他部分太無效了所以完全被掩蓋掉了也說不一定．

＊How to improve your performance

因為任務還沒成功，所以也沒有甚麼performance可以outstanding，有測試過像投影片一樣把字典數量降低，只要只出現一次的字就全部改用一個"＜HARD＞"代替，不過效果不明顯．

有測試過雙向或單向LSTM，雙向更好一點，有嘗試過把FEATURE倒過來輸入，反正對機器來說全部反過來是一樣的吧？應該沒有不合文法的問題...不過預測準確力還是不好，總之還是只會A man... A man... A man...．

另外測試過的是將８０個time step的輸出做完後將隱藏層狀態存起來，然後再做為下一個ＬＳＴＭ的初始狀態，而第二個ＬＳＴＭ的輸入長度就是等同於輸出time step的長度，只是輸入的特徵用全零然後都mask掉，也就是不使用Repeat Vector的方式而純粹用encoder完後的狀態來輸出，這樣子也是不work，loss都無限大．

＊Experimental results and settings

在實際測試的時候都是採用每一輪在1450中，每種都隨機挑出一個LABEL，順序是按照FEAT的順序每次都是1~1450，訓練的BATCH SIZE是設定25．

大概只要在第三輪之後就不會再上升了，然後通常在第10次感覺準確率就有點變低，

都是在接近0.2左右・

由於挑選出來的句子都介在７～１２之間，平均以１０來看，這樣表示每次只猜對差不多２個字，當然就是A man... A man... A man...，非常地令人感到哀傷＜EOS＞